Robust Open Vocabulary Translation from Visual Text Representations

Elizabeth Salesky, David Etter, Matt Post

January 17, 2024

Abstract

This is a paper review for Robust Open Vocabulary Translation from Visual Text Representations.

1 Summary and contributions

As machine translation degrades quickly due to presence of noise, and researchers are in the position of synthetically adding noisy training data, Elizabeth Salesky, David Etter and Matt Post present a different idea for Machine Translation.

The authors propose a translation from visual text representation instead of using the classic methods of text analysis, arguing that humans process text visually more than unicode representations.

In the experimental setup it is presented to us that two data scenarios were used in this experiment. A small one Multitarget TED Talks Task (MTTT) of aproximately 200k training sentences, and a larger one Conference on Machine Translation (WMT20).

In 3.2 it is being mentioned that test sets used were created from Reddit comments in French, German and Japanese which have been professionally translated in English. We also learn that authors used the WIPO corpus to evaluate the occurring noise for Russian-English.

Authors introduces synthetic noise to make a comparison between visually similar characters and character permutations in Cambridge.

Phenomena	Word	BPE (5k)	
Vowelization	كتاب	كتاب	(1)
	الكِتابُ	ال ـ ِ ـ ت ـ اب ـ ُ	(5)
Misspelling	language	language	(1)
	langauge	la ـ ng ـ au ـ ge	(4)
Visually Similar Characters	really	really	(1)
	rea11y	re ـ a ـ 1 ـ 1 ـ y	(5)
Shared Character Components	확인한다	확인 ـ 한 ـ 다	(3)
	확인했다	확인 ـ 했다	(2)

Figure 1: Examples of common behavior which cause divergent representations for subword models [1].

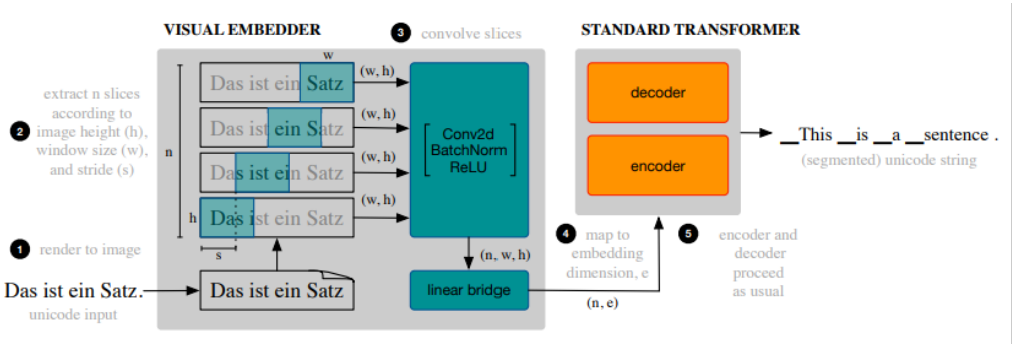

Figure 2: Visual text architecture combines network components from OCR and NMT, trained end to end [1]

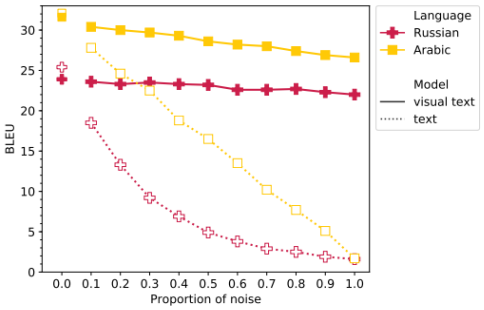

Figure 3: Visual noise [2]

They have tested noises as follows: unicode, diacrticitcs, 133tspoke, swap, cambridge

As Cambridge spelling experts illustrates robustness of human to character permutation when first and last letters are the same, noise can be applied to words with length ≥ 4 Ḣence, authors did not apply it to Chinese and Japanese texts, as most words contain fewer characters after word segmentation.

1.1 Conclusions

- First conclusion in 5. *"We find font size is not significant as long as it is sufficiently large to not affect image resolution for more visually dense scripts" [2]*
- *"Character based models are similarly unable to handle OOV codepoints, and characters in extremely novel contexts, as found with this type of noise: at $p = 0.5$, our character model has a disappointing 0.2 BLEU [2]"*
- *"Inducing visually similar codepoint differences barely affects visual text, but breaks BPE representations" [2] .*
- *"while both visual text models and text models are negatively affected by induced l33tspeak, the visual text models for both language pairs significantly outperform the text models in these conditions." [1]*

- *"Character permutations are challenging both for subword models, which necessarily back off to smaller units in the presence of OOVs (Table 1), and character based models" [1]*
- Visual text representations result in significant improvements for character permutations, particularly at higher levels of noise
- Authors introduces visually rendered text for continuous open-vocabulary translation, trained on seven pairs and two data settings, approach or matched the performance of traditional text models
- Authors approach operates on raw text, so they do not need to include normalization, tokenization and subword segmentation

2 Strenghts

- In introduction the authors explain clearly the aim of the paper
- Authors present and motivate their choices
- Authors explain how their architecture works, provide a graphical representation of it, and compare their approach to the traditional one.
- Authors described testing two data scenarios on experimental setup.
- Evaluation model is clearly explained at 3.2
- Authors included related work
- Authors have solid conclusions
- Future work was included as part of the paper
- Clear and correct references

3 Weaknesses

- We learn quite late about the languages between which translation is made. (3.2)

- In 3.3 we find out that they used 7 language pairs from MTTT TED, before we know what are the languages.
- Authors clearly present the packages being used to render text for each language.
- Authors acknowledge that, while effective, it is not clear that sliding window segmentation is optima

4 Correctness

- The datasets chosen for the NLP paper were sourced from reputable repositories

5 Clarity

- Aim of the paper is clear from the beginning
- Clarity on how data set was chosen
- Authors explained how they induces noise, and the resulting inputs and outputs for the text and visual text models, on each type of tested noise.
- It is clear how they conducted the study, they included annexes with parameter tuning information, experiments

6 Additional Feedback

As an additional feedback, I would like to mention that it impresses me how authors included detailed explanations of how they conducted the study, steps on how to reproduce it and a separate GitHub thread on their work.

7 Overall Score

I liked this article as it presents a creative idea on vocabulary translation, and I would rate it with an overall score of 4. It provides a creative idea

on vocabulary translation. Although it is not clear that sliding window segmentation is optimal, and that there is clearly the opportunity for further improvement, the authors did a great job in clearly explaining their work. The paper is comprehensive, and it contains all relevant information of a scientific paper.

Grade Criterion:

Comprehensibility (incomprehensible = 1, easy to understand = 5) 5
Information density (very low = 1, very high = 5) 3
Content (many content errors = 1, consistent content = 5) 5
Accuracy (very inaccurate = 1, very accurate = 5) 5
Key message (not recognizable = 1, very clear = 5) 5
Abstract (misses the point = 1, very appropriate = 5) 4
Summary (misses the point = 1, very appropriate = 5) 4
References (not available = 1, very appropriate = 5) 5
Drawings, tables (not available = 1, clear and concise = 5) 5
Length of the paper (far off = 1, within specification = 5) 5

References

[1] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[2] Elizabeth Salesky, David Etter, and Matt Post. Robust open-vocabulary translation from visual text representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November 2021. Association for Computational Linguistics.